Introduction to Chat GPT

Chat GPT is a part of the GPT i.e. Generative Pre trained Transformer which transform the wide amount of text data into generate human like text. Chat GPT is one of the large language models that can perform NLP (Natural Language Processing) tasks such as text summarization, question answering , machine translation grammatically error correction within a single architecture. This model has demonstrated its potential in different sectors such as machine-human interaction, research, education, business, health and reasoning. It is a successor to Instruct GPT, a framework based on RLHF (Reinforcement Learning from Human Feedback) that allows the model to set with human values and preferences. It improves from LLMs (Large Language models) which are trained text corpora through unsupervised training. Now Chat GPT has gained remarkable attention and interest from many applicants due to its natural response and exceeding potential. (Tu, 6 Apr 202) .

The main objective of Chat GPT is to make the interaction between human and artificial intelligence more natural. Chat GPT provides more advancement over traditional bots found in the market.

History background of Chat GPT

In the computer science community, Artificial Intelligence Generated Content (AIGC) has gained, beyond attention where Chat GPT and DALL-E 2 were introduced by large tech companies. AIGC is the given human instructions using GAI algorithms () that helps to guide and teach the model to complete the task.

Recent AIGC has more sophisticated models for large datasets that uses larger foundation model architectures and accesses to extensive computational resources compared to prior works. Similarly both Chat GPT and DALLE-E 2 were developed by Open AI which can understand and respond to human language and are able to create high quality and unique images from textual descriptions within a few minutes respectively. The framework of GPT-3 maintains the same as GPT-2 but has better generalization ability than GPT-2 in terms of various tasks such as human intent extraction. The pre training data sizes grow from Web text (38GB) to Common Crawl (570GB after filtering) and the foundation model size grows from 1.5B to 175B.[ 8 and 9.] New technologies were being developed with GAI algorithms due to increase in computational power and data volume that brought many benefits. There are three types of machine learning i.e. supervised learning, unsupervised learning and reinforcement learning.

Open AI initiative was founded by Sam Altman, El’ Musk and others. At the forefront of AI research, he has produced several breakthrough models such as GPT-2, GPT-3 and finally Chat GPT building on the success of GPT-3. Open AI continued to lead the creation with R&D efforts and Chat GPT based on GPT-4 Architecture was introduced.

The evolution of GPT

GPT – 1: This is the first version of GPT released on 2018. A neural network architecture i.e. Transformer architecture was used in this version to do NLP (Natural Language Processing) tasks such as machine translation and language modeling. Its architecture is of 12 level, 12 – headed transformer decoder (no encoder), followed by linear-soft max with Book Corpus: 4.5 GB of Text. It has 117 million of parameter count.

GPT - 2: It is the improvised version of GPT -1 with 1.5 billion parameters with greatest parameters at that time. It was released on 2019 with modified normalization with Web Text: 40 GB of text. The notable feature of this version was it can generate coherent and realistic text which is difficult to differentiate from the human-written text. This raised some concerns about possible misuse of this model, such as generation of fake news. This is why Open AI initially decided not to release a full version but released smaller version.

GPT-3: This was released on 2020 with the modification to allow larger scaling with 570 GB plaintext. With 172 billion parameters, it was the most powerful and largest model ever created. This model’s ability is to perform sentiment analysis, question and answering of NLP. It also includes language translations, chat bots, code generations etc. It has sparked new research and development in the field of open AI.

Other GPTs:

Instruct GPT, ProtGPT2, Bio GPT were other GPTs that were introduced after GPT-3 in 2022 along with Chat GPT. Instruct GPT has same parameter count as GPT-3 but is fine tuned to follow instructions using human feedback model where ProtGPT2, Bio GPT has 738 and 347 million of parameters count respectively. ProtGPT2 is built on the GPT2 Transformer architecture and includes 36 layers with a model dimensionality of 1280, making it a powerful model with 738 million parameters. The pre-training of ProtGPT2 was done on the UniRef50 database (version 2021\_04) in a self-supervised manner, using raw protein sequences without any annotation. Whereas Bio GPT is based on TLM (Transfer model architecture) focused on mining and generating biomedical text. And Chat GPT with 175 billion parameters count uses version GPT-3.5 with fine tuned of both reinforcement learning from human feedback and supervised learning.

GPT- 4:

It is the latest version of GPT with 100 trillion of parameter count which is trained with both text prediction, RLHF and accepts both text and images as input as well as from third party. The development of GPT-4 took advantage of lessons learned from the Open AI and Chat GPT adversarial testing programs and made iterative adjustments over a period of six months, resulting in improvements in terms of factuality, controllability, and boundary compliance. Still there is room for improvement.

How Chat GPT works:

The working procedure of Chat GPT can be divided into two types i.e. Query procedure and Response.

An artificially intelligent super computer is the device behind the Chat GPT. These computers are trained with numerous parameters on a massive data set which is unsupervised learned by determining the statistical structure within the data to identify the patterns. Usually, user makes a Query to the Chat GPT. Then that query is sent to the super computer and is processed. The probable output is generated by the query generation circuitry and that output data are fine tuned. After this process, Chat GPT is directed to response. Finally, the conversational interface interacts with human and provide human like response in Chat GPT.

Since Chat GPT utilizes the GPT-3.5. The followings are the workflow of GPT 3.5:

1. Collect demonstration data and train a supervised strategy. First, prompts are sampled from the prompt record (dataset). Then demonstrated by labeler with desired output behavior. This data will be used to fine-tune his GPT3 in supervised learning.
2. Collecting comparative data and training reward models. Then the prompt and some model output are sampled. Labeler organizes the output from best to worst.
3. Policy optimization for reward models using reinforcement learning.

Finally, new prompts are sampled from the dataset. The policy produces output of a reward model and computer rewards for outputs.